Approximations on Risk-Averse Markov Decision Processes

Eugenio Della Vecchia, Silvia C. Di Marco, Alain Jean-Marie

Approximations on Risk-Averse Markov Decision Processes

Eugenio Della Vecchia , Silvia Di Marco , Alain Jean-Marie

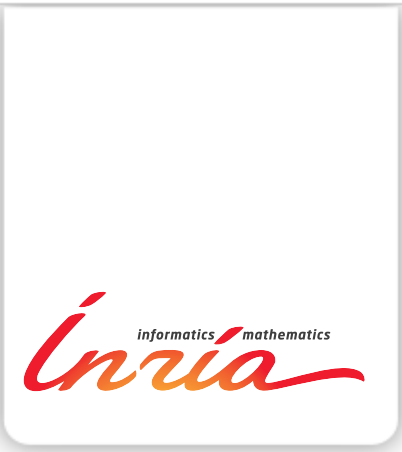
RESEARCH REPORT

N° 8393

November 2013

Project-Team Maestro

ISSN 0249-6399 ISRN INRIA/RR--8393--FR+ENG



Approximations on Risk-Averse Markov Decision Processes

Eugenio Della Vecchia *, Silvia Di Marco †, Alain Jean-Marie‡

Project-Team Maestro

Research Report n° 8393 — November 2013 — 20 pages

Abstract: We consider the problem of approximating the values and the optimal policies in risk-averse discounted Markov Decision Processes with infinite horizon.
We study the properties of the rolling horizon and the approximate rolling horizon procedures, proving bounds which imply the convergence of the procedures when the horizon length tends to infinity.
We also analyze the effects of uncertainties on the transition probabilities, the cost functions and the discount factors.

Key-words: Markov Decision Processes, Rolling Horizon, risk aversion

* CONICET - UNR, Argentina
† CONICET - UNR, Argentina
‡ INRIA and LIRMM, CNRS/Université Montpellier 2, 161 Rue Ada, F-34392 Montpellier, ajm@lirmm.fr.

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Approximations dans les processus de décision Markoviens averses au risque

Résumé : Nous considérons le problème de l'approximation de la fonction de valeur et des politiques optimales dans un processus de décision Markovien avec actualisation et aversion au risque.

Nous étudions les propriétés de la procédure de l'horizon roulant et son approximation, et montrons des bornes qui impliquent la convergence de ces procédures quand l'horizon de temps tend vers l'infini.

Nous analysons aussi les effets d'incertitudes sur les probabilités de transition, les fonctions de coût et les facteurs d'actualisation.

Mots-clés : Markov Decision Processes, horizon roulant, aversion au risque

1 Introduction

Consider a random dynamical system, observed at discrete times. At each time $t \in \mathbb{N}$, the state s_t is observed, an action a_t is chosen, and the system moves to a new state s_{t+1}, resulting in an instantaneous cost (or negative reward) $r^{a_t}(s_t, s_{t+1})$. The cost function r is supposed to be independent on time. The actions a_t chosen at each time t in the respective state s_t determine a policy π whose performance is evaluated through a *dynamic risk measure*. Dynamic risk measures generalize the expected discounted sums of all costs, to which they reduce in the "risk neutral" case.

The objective of the controller is to find (when it exists) the policy that minimizes, given the current state s, the total performance defined over the infinite time horizon. However, for a wide class of stochastic control problems in discrete time and infinite horizon, obtaining an optimal policy explicitly is a difficult task. This is why practitioners often use instead a heuristic method called the Rolling Horizon (**RH**) procedure

On the other hand, in the context of stochastic sequential decision models, the controller has to take decisions, based on the knowledge of the current state, but without the certainty of the dynamics of the system, which will be governed by distributions of probabilities on the space of states, known *a priori*.

In both cases, an approximation of some original problem is solved. In such situations, it is useful to know bounds on the error incurred by solving approximated problems. The robustness of the model can indeed be assessed by analyzing the sensitivity of the optimal controls and the value function when considering approximations of the elements of the models.

This analysis has been the topic of many publications in the case where the "performance" of the system is measured with the expected total discounted cost. For instance in [8] in the **MDP** (Markov Decision Process) context and in [5] for Semi-Markov Games (and consequently for Semi-Markov Decision Problems). Structural approximations for Semi-Markov Games are obtained in [6].

On the other hand, using the expected cost as a metric is known to be characteristic of "risk-neutral" optimizing agents. One may want to use a different, risk-sensitive metric: this is the purpose of risk measures. The use of dynamic risk measures in the context of optimal control problems of the **MDP** type, was proposed by A. Ruszczyński in [12] and in a posterior work with O. Çavuş in [3]. The purpose of the present work is to analyze the Rolling Horizon heuristic and perform a sensitivity analysis in this new context. We therefore generalize the results of [5, 6] to dynamic risk measures. The approximations obtained for **MDPs** in the literature are then recovered by specializing to the risk-neutral case. In [10], the authors use the terminology *dynamic risk mappings*, and in [11, 12, 3], they talk about *dynamic risk measures*. In this work we adopt this last terminology.

The paper is organized as follows. We complete this introduction with a brief literature review of the approximation techniques made in the risk neutral case, similar to the ones proposed in this work. In Section 2 we make more precise the elements of the **MDP** model, following [7], and introduce the definitions and results on dynamic risk measures following [12] and [3]. Then in Sections 3 and 4 we state the **RH** procedures and the structural approximations on the elements of the models, respectively, and give bounds for the errors incurred in each case. An example developed in Section 5 illustrates some of these bounds and raises some issues. Conclusions are presented in Section 6.

Rolling Horizon Approximations are used in [8] to obtain almost optimal stationary policies in discounted and average **MDP** models. In [4], the authors analyze the performance of the procedures when they are applied on Markov Games (**MG**). These authors also introduce the

idea of approximate Rolling Horizon. On the other hand, structural approximations were studied in the book [1], on constrained **MDP**, in [13], for the **MG** case, and in [6], on Semi-Markov Games.

2 Preliminaries and Notations

2.1 Markov Decision Processes

We consider a Markov Decision Process of the form

$$\mathcal{M} := (\mathcal{S}, \mathcal{A}, \{\mathcal{A}_s : s \in \mathcal{S}\}, \{Q^a(\cdot|s) : s \in \mathcal{S}, a \in \mathcal{A}_s\}, r, \alpha) \ , \tag{1}$$

where $\mathcal{S}$ is the state space; for every $s \in \mathcal{S}$, $\mathcal{A}_s$ is the set of actions available in state s and $\mathcal{A} = \bigcup_{s\in\mathcal{S}} \mathcal{A}_s$ is the action space. We set $\mathbb{K} = \{(s, a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$. Given a state s and an action $a \in \mathcal{A}_s$, the transition law $Q^a(\cdot|s)$ is a stochastic kernel on $\mathcal{S}$. The costs of transitions between states is a function $r : \mathbb{K} \times \mathcal{S} \mapsto \mathbb{R}$ and α is a discount factor.

For Borel sets X and Y, we will note with $\mathbb{P}(X)$ the family of probability measures on X endowed with the weak topology, and with $\mathbb{P}(X|Y)$ the family of transition probabilities from state Y to X.

The space $H_t := \mathbb{K}^t \times \mathcal{S}$ of admissible histories of the process at the t-th decision epoch, consists of sequences of states and decisions up to that epoch.

A Markov policy is a sequence $\pi = \{\pi_t\}$ of stochastic kernels $\pi_t \in \mathbb{P}(\mathcal{A}|H_t)$ such that for every $h_t \in H_t$ and $t \in \mathbb{N}$, $\pi_t(\mathcal{A}_{s_t}|h_t) = 1$. We denote by Π the set of all Markov policies. A Markov policy $\pi = \{\pi_t\}$ is stationary if there exists $f \in \mathbb{P}(\mathcal{A}|\mathcal{S})$ such that $f(s) \in \mathbb{P}(\mathcal{A}_s)$ and $\pi_t = f$ for all $s \in \mathcal{S}$ and $t \in \mathbb{N}$. In this case, we identify π with f, i.e., $\pi = f = \{f, f, ...\}$. We denote by Π_{stat} the set of all stationary policies.

For each policy $\pi \in \Pi$ and any initial state s there exist a unique probability measure P_s^π and stochastic processes $\{S_t\}$ and $\{A_t\}$. S_t and A_t represent the state and the actions at the t-th decision epoch. $\mathbb{E}_s^\pi$ denotes the expectation operator with respect to P_s^π.

In general, for functions defined on $\mathcal{S}$, we shall not write the state variable s, but sometimes it will be convenient to emphasize this dependency. In addition, given a stationary policy $f \in \Pi_{\text{stat}}$ we let $\phi^f(s) = \phi^{f(s)}(s)$. The notation is also extended to probability distributions: for ξ on $\mathcal{A}_s$, $h^\xi(s) = \int_{\mathcal{A}_s} h^a(s)\xi(\mathrm{d}a)$ whenever the integral is well defined. In the same spirit, we shall denote $r^a(s, s') = r(s, a, s')$.

In order to evaluate the performance of policies, a discounted criterion is often used. For $N \geqq 1$, $s \in \mathcal{S}$, $\pi \in \Pi$, the expected N-horizon α-discounted cost is defined by

$$V_N^\pi(s) = \mathbb{E}_s^\pi \left[\sum_{t=0}^{N-1} \alpha^t r^{A_t}(S_t, S_{t+1})\right] \ ,$$

and for the infinite horizon problem by

$$V^\pi(s) = \mathbb{E}_s^\pi \left[\sum_{t=0}^{\infty} \alpha^t r^{A_t}(S_t, S_{t+1})\right] \ .$$

In this so called risk neutral case, and under standard assumptions, the problem has Markov optimal policies for finite-horizon models and stationary optimal policies in the finite-horizon case.

2.2 Introduction of Markov Risk Measures in the MDP

In [12] A. Ruszczyński introduces risk aversion to dynamic problems, replacing the expected value operator by a dynamic risk measure. In [3] the use of dynamic risk measures is extended to transient problems, in which there exists an state s^* such that for all $a \in A_{s^*}$, $Q^a(s^*|s^*) = 1$ and $r^a(s^*, s^*) = 0$, sometimes referred to as "stochastic shortest path problems". We adopt from those works the notations and refer to them for the definitions involved.

We assume at first that a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is fixed. $\mathcal{F}_t$ will denote the σ-subalgebras generated by the histories h_t, and define spaces $\mathcal{Z}_t$ of $\mathcal{F}_t$-measurable random variables on Ω. Precisely, we shall work with $\mathcal{Z}_t = L^p(\Omega, \mathcal{F}_t, \mathrm{P})$, for some $p \in [1, \infty)$, the spaces of real-valued, $\mathcal{F}_t$-measurable and p-integrable random variables on Ω. For details on the construction and generalizations on the fixed probability space we refer to [3, Section 3].

A Markov risk measure ([12, Definition 6, p. 245]) with respect to the process $\{S_t\}$ is a one-step conditional risk measure $\rho_t : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$ ([12, Definition 1, p. 239]) such that there exists a risk transition mapping $\sigma(v, s, Q)$ ([12, Definition 5, p. 243] or [3, Definition 3.1, p. 3]) that satisfies, for any bounded function v,

$$\rho_t(v(S_{t+1})) = \sigma(v, s_t, Q^{a_t}(\cdot|s_t)) \ .$$

The r.h.s. of this expression is parametrized by the state s_t, and it defines a function $\mathcal{S} \to \mathbb{R}$.

Suppose that N is a fixed horizon. Each policy $\pi = \{f_0, f_1, ..., f_{N-1}\}$ results in a cost sequence $Z_t = r^{A_{t-1}}(S_{t-1}, S_t)$, $t = 1, ..., N$. The risk of this sequence in the finite-horizon case will be evaluated by using the functional [3, Equation (1), p. 3]

$$\begin{aligned} J_N^\pi(s) \quad = \quad & \rho_1\big(r^{A_0}(s, S_1) + \alpha\rho_2\big(r^{A_1}(S_1, S_2) \\ & +\alpha^2\rho_3(r^{A_2}(S_2, S_3) + ... + \alpha^{N-1}\rho_N(r^{A_{N-1}}(S_{N-1}, S_N))...)\big)\big) \ , \end{aligned}$$

where ρ_t are the Markov conditional risk measures introduced above.

For the infinite horizon context the risk will be evaluated using discounted risk functional [12, Sections 6-7, pp. 248–250] of the form

$$J^\pi(s) = \lim_{N\to\infty} J_N^\pi(s) \ . \tag{2}$$

In order to prove optimality results, the following dynamic programming operators on the space of bounded functions are introduced. For any stationary policy $f \in \Pi_{\text{stat}}$, and $s \in \mathcal{S}$,

$$(T^f v)(s) = \sigma\left(r^f(s, \cdot) + \alpha v(\cdot), s, Q^f(\cdot|s)\right) \tag{3}$$

and

$$(Tv)(s) = \inf_{a\in\mathcal{A}_s} \sigma\left(r^a(s, \cdot) + \alpha v(\cdot), s, Q^a(\cdot|s)\right) \ , \tag{4}$$

where, for $s \in \mathcal{S}$, $\sigma(\cdot, s, Q)$ is the risk transition mapping associated to the probability distribution $Q \in \mathbb{P}(\mathcal{S}|\{s\})$ on $\mathcal{S}$ given s.

As it is also developed in [12, Section 4, p. 243–246], if for any $s \in \mathcal{S}$ and $Q \in \mathbb{P}(\mathcal{S}|\{s\})$, σ is lower semicontinuous with respect to the first argument, then there exists a closed convex multifunction $\mathfrak{A}(s, Q)$, with values in $\mathbb{P}(\mathcal{S})$, such that for every bounded real function v,

$$\sigma(v, s, Q) = \sup_{m\in\mathfrak{A}(s,Q)} \int_{\mathcal{S}} v(z)\mathrm{d}m(z) \ . \tag{5}$$

We can now state our working assumptions and some preliminary results.

Assumption 1.

(a) *$\mathcal{S}$ is a Borel set.*

(b) *For each $s \in \mathcal{S}$, $\mathcal{A}_s$ is a compact set.*

(c) *For each $s \in \mathcal{S}$ the mapping $a \mapsto Q^a(\cdot|s)$ is continuous in $\mathcal{A}_s$.*

(d) *For each $t \in \mathbb{N}$, the conditional risk measure ρ_t is a Markov risk measure, and such that for every $s \in \mathcal{S}$, $\mathfrak{A}(s,\cdot)$ is lower semicontinuous.*

(e) *r is a bounded function on $\mathbb{K} \times \mathcal{S}$: there exists $M \in \mathbb{R}$ such that $||r||_\infty = M$.*

(f) *For each $s, s' \in \mathcal{S}$, $r^a(s,s')$ is a lower semicontinuous function on $\mathcal{A}_s$.*

Lemma 2.1. *If* **Assumption 1** *holds, for any stationary policy $f \in \Pi_{\text{stat}}$, T^f in a non decreasing and α-contractive operator on the space of bounded functions. The same properties hold for the operator T.*

Moreover, the value J^f of a stationary policy f is the unique bounded solution of the equation $T^f v = v$.

Proof. See [12, Lemma 1, p. 252], [12, Lemma 2, p. 252] and [12, Lemma 4, p. 253]. □

Theorem 2.1. *Suppose that* **Assumption 1** *holds. Then*

(a) *For all $\pi \in \Pi$, $||J^\pi||_\infty \leqq \frac{M}{1-\alpha}$.*

(b) *The finite horizon problems have optimal values. Moreover, starting from $J_0^* \equiv 0$, for $n \geqq 1$, the function $J_n^* := TJ_{n-1}^*$ is the value function for the n-horizon problem, and the Markovian policy $\{f_{n-1}^*, f_{n-2}^*, ..., f_1^*, f_0^*\}$, where the functions f_k^* are the corresponding minimizing functions, for $k = 0, ..., n-1$, is optimal.*

(c) *The infinite-horizon problem has a value J^*, and for all $s \in \mathcal{S}$,*

$$|J^*(s) - J_n^*(s)| \leqq \frac{M\alpha^n}{1-\alpha} \to 0$$

as $n \to \infty$.

(d) *J^* is the unique bounded function satisfying the optimality equation $Tv = v$.*

Moreover, there exists an stationary policy f^* which is optimal for the infinite-horizon problem.

Proof. See [12, Theorem 2, p. 246], [12, Theorem 4, p. 250] and [12, Theorem 5, p. 254]. □

3 Rolling Horizon Approximation Procedures

3.1 The Rolling Horizon Procedure

When large stochastic control problems are analyzed, the complete computation of explicit optimal strategies is a difficult, or even an impossible task. Instead, in practice a heuristic method called the Rolling Horizon procedure (also, Receding Horizon, or Model Predictive Control) is often used. The **RH** method prescribes to repeatedly solve a finite-stage horizon problem, taking the current state as initial state. Then, only the first decision will be applied.

Specifically, the procedure to construct a **RH** policy is the following one. Fix some integer N (the horizon length) and consider a sequence of epochs indexed by $t \in \mathbb{N}$.

RH1 At iteration t, and for the current state s_t, solve the finite horizon problem (**FHP**) with horizon N. In the context of risk measures, this problem would be:

$$(\mathbf{FHP}) \qquad \inf_{\pi} \rho_1\big(r^{A_0}(s, S_1) + \alpha\rho_2\big(r^{A_1}(S_1, S_2) \\ +\alpha^2\rho_3(r^{A_2}(S_2, S_3) + ... + \alpha^{N-1}\rho_N(r^{A_{N-1}}(S_{N-1}, S_N))...)\big)\big) \ ,$$

taking $s = s_t$ as initial state. An action $f_{N-1}(s_t)$ is obtained.

RH2 Apply $a_t = f_{N-1}(s_t)$.

RH3 Observe the achieved state at time $t+1$: s_{t+1}.

RH4 Set $s_t := s_{t+1}$ and $t := t+1$ and go to step **RH1**.

The **RH** procedure does not specify how to compute the value $f_{N-1}(s_t)$. Its efficiency is based on the idea that computing the value $f_{N-1}(s_t)$ alone is usually much easier than computing the N decision rules $(f_{N-1}, f_{N-2}, \ldots, f_1, f_0)$ that are the usual result of solving **(FHP)** for all possible initial states. On the other hand, the performance of the resulting sequence of decisions is not the optimal one, although the intuition is that when N is "large enough", the performance should be close to the optimal. The practical issue is then to choose N so as to obtain a proper compromise between precision and the computational effort needed to obtain $f_{N-1}(s_t)$. We address this issue through two formal qualitative and quantitative questions. Let $U_N(s)$ be the cost incurred by using the **RH** policy of horizon length N, starting in state s:

Q1 Under which conditions on the problem is it true that $\lim_{N\to\infty} U_N(s) = J^*(s)$?

Q2 Given a state s and $\varepsilon > 0$, is it possible to compute N such that $U_N(s) - J^*(s) < \varepsilon$?

In what follows we prove the convergence of the procedure to the value of the original problem, stated as Theorem 3.1 and Theorem 3.2. The term "convergence" has to be understood in the sense that when the horizon N goes to infinity, the value obtained with the procedure approaches the value of the problem as in **Q1**. The preliminary observation, classical for studying **RH**, is that the procedure effectively implements a stationary policy. Since the controller will repeatedly act according to the state-feedback function f_{N-1}, we have $U_N = J^{f_{N-1}}$.

In order to improve the bounds in the next results, we consider the following additional assumption, which refers to the sequence J_n^* of Theorem 2.1.

Assumption 2. *For all $s \in \mathcal{S}$, $J_1^*(s) \geqq 0$.*

Observe that this is the case, for instance, when the cost function r is positive. It is easy to prove that if **Assumptions 1** and **2** hold, then for $n \in \mathbb{N}$ and $s \in \mathcal{S}$, $J_n^*(s) \leqq J_{n+1}^*(s)$.

Theorem 3.1. *Suppose that* **Assumption 1** *hold. Then, for all $s \in \mathcal{S}$,*

$$0 \leqq U_N(s) - J^*(s) \leqq \frac{2M\alpha^N}{1-\alpha} \ ,$$

and in consequence

$$||J^* - U_N||_\infty \leqq \frac{2M\alpha^N}{1-\alpha} \ .$$

If in addition **Assumption 2** *holds,*

$$||J^* - \tilde{U}_N||_\infty \leqq \frac{M\alpha^N}{1-\alpha} \ .$$

Proof. This result is a corollary of Theorem 3.3, taking $\varepsilon = 0$. Indeed, in that case $J = J^*_{N-1}$ and $T^f J = T J^*_{N-1} = J^*_N$, which means that $\tilde{f}_{N-1} = f_{N-1}$ and $\tilde{U}_N = U_N$. □

Theorem 3.2. *Let J^*_n be the sequence of functions defined in Theorem 2.1, part* (b)*. Then, if $||J^*_N - J^*_{N-1}||_\infty \leqq \varepsilon$, then the* **RH** *policy f_N is $\frac{2\alpha\varepsilon}{1-\alpha}$-optimal. That is,*

$$||J^* - U_N||_\infty \leqq \frac{2\alpha\varepsilon}{1-\alpha} .$$

Proof. This result is a corollary of Theorem 3.4, taking $\varepsilon_2 = 0$. □

3.2 An Approximate Rolling Horizon Procedure

Suppose now that the controller does not have exact information about the problem to be solved at **RH1** in the **RH** procedure, but he knows or he is able to compute an approximation of that value. We are interested in implementing a procedure where this last approximation is used instead of the value function of the problem with finite horizon and estimate the error introduced.

Then, for a function J, supposed to be close in some sense to J^*_{N-1}, choose

$$\tilde{f}_N(s) \in \arg\min_{a \in \mathcal{A}_s} \sigma\left(r^a(s,\cdot) + \alpha J(\cdot), s, Q^a(\cdot|s)\right) .$$

Specifically,

ARH1 Choose some function J a priori near J^*_{N-1} where J^*_{N-1} is the $N-1$-stage value function.

ARH2 At iteration t, and for the current state s_t, solve

$$\min_{a \in \mathcal{A}_{s_t}} \sigma\left(r^a(s,\cdot) + \alpha J(\cdot), s_t, Q^a(\cdot|s_t)\right).$$

An action $\tilde{f}_{N-1}(s_t)$ is obtained.

ARH3 Apply $a_t = \tilde{f}_{N-1}(s_t)$.

ARH4 Observe the achieved state at time $t+1$: s_{t+1}.

ARH5 Set $s_t := s_{t+1}$ and $t := t+1$ and go to step **ARH2**.

We will note with $\tilde{U}_N$ the total discounted reward of the stationary policy $\tilde{f}_{N-1}$. Theorem 3.3 gives answers to questions **Q1** and **Q2** stated in this section for the sequence of successive rewards $\tilde{U}_N$. We begin with a lemma of general character.

Lemma 3.1. *Suppose that* **Assumption 1** *holds. Let $f \in \Pi_{\text{stat}}$ be a stationary policy, with total discounted cost J^f, v a bounded function and C a positive constant such that*

$$v \geqq T^f v - C ,$$

then, for any $n \in \mathbb{N}$,

$$(T^f)^n v \geqq J^f - \frac{C\alpha^n}{1-\alpha} .$$

Proof. First of all, observe that:

$$\begin{aligned} T^f(v-C) &= \sigma\left(r^f + \alpha(v-C), \cdot, Q^f\right) \\ &= \sigma\left(r^f + \alpha v - \alpha C, \cdot, Q^f\right) \\ &= \sigma\left(r^f + \alpha v, \cdot, Q^f\right) - \alpha C \\ &= T^f v - \alpha C \ . \end{aligned}$$

The third equality is an easy consequence of (5). Then, by the monotonicity of T^f, we have successively:

$$\begin{aligned} v &\geqq T^f v - C \\ T^f v &\geqq T^f(T^f v - C) = (T^f)^2 v - \alpha C \end{aligned}$$

and by recurrence, it is clear that for all t:

$$(T^f)^t v \ \geqq \ (T^f)^{t+1} v - \alpha^t C \ .$$

Summing up these inequalities for t from n to m, we obtain:

$$(T^f)^n v \ \geqq \ (T^f)^{m+1} v - C \sum_{t=n}^{m} \alpha^t \ . \tag{6}$$

Since T^f is contractive with J^f as its unique fixed point (Lemma 2.1), for any bounded function v,

$$\lim_{m\to\infty} (T^f)^{m+1} v = J^f \ .$$

Finally, the second term in the right-hand side of (6) converges, as $m \to \infty$, to $\frac{C\alpha^n}{1-\alpha}$, and the stated bound follows.

□

Theorem 3.3. *Suppose that* **Assumption 1** *holds. Given J a bounded function such that for some $N \geqq 0$, $||J^*_{N-1} - J||_\infty \leqq \varepsilon$, consider a policy $f \in \Pi_{\text{stat}}$ such that $T^f J = TJ$. Then,*

$$0 \leqq \tilde{U}_N(s) - J^*(s) \leqq \frac{2M\alpha^N}{1-\alpha} + \frac{2\alpha\varepsilon}{1-\alpha} \ ,$$

and in consequence

$$||J^* - \tilde{U}_N||_\infty \leqq \frac{2M\alpha^N}{1-\alpha} + \frac{2\alpha\varepsilon}{1-\alpha} \ .$$

If in addition **Assumption 2** *holds,*

$$||J^* - \tilde{U}_N||_\infty \leqq \frac{M\alpha^N}{1-\alpha} + \frac{2\alpha\varepsilon}{1-\alpha} \ .$$

Proof. As a prelude to the proof, let us first note that if r is bounded by M, the random variables $Z_t = r^{A_{t-1}}(S_{t-1}, S_t)$, which represents the gain at time $t-1$ by application of the Markov policy π, satisfy $-M \leqq Z_N \leqq M$, and following the ideas in the proof of [12, Theorem 3, pp. 249-250] the next inequality holds:

$$\rho^\alpha_{1,N}(Z_1, ..., Z_N) \leqq \rho^\alpha_{1,N-1}(Z_1, ..., Z_{N-1}) + M\alpha^{N-1} \ ,$$

with the notation, for $t \in \mathbb{N}$,

$$\rho^{\alpha}_{1,t}(Z_1, ..., Z_t) = \rho_1 \left(Z_1 + \alpha \rho_2 \left(Z_2 + \alpha^2 \rho_3 \big(Z_3 + ... + \alpha^{t-1} \rho_{t-1}(Z_t) \big) ... \right) \right) \ .$$

Then, for all $N \in \mathbb{N}$ the value functions of the problems with horizons $N-1$ and N verify

$$\begin{aligned} J^*_{N-1} &= \inf_{\pi} \ \rho^{\alpha}_{1,N-1}(Z_1, ..., Z_{N-1}) \\ &\geqq \inf_{\pi} \ \rho^{\alpha}_{1,N}(Z_1, ..., Z_N) - M\alpha^{N-1} \\ &= J^*_N - M\alpha^{N-1} \ . \end{aligned} \tag{7}$$

Let us start the proof using the triangular inequality

$$||J^* - \tilde{U}_N||_\infty \leqq ||J^* - TJ||_\infty + ||TJ - \tilde{U}_N||_\infty \ . \tag{8}$$

To bound the first term in the right-hand side of (8), let us observe that, by Lemma 2.1,

$$||TJ^*_{N-1} - TJ||_\infty \leqq \alpha ||J^*_{N-1} - J||_\infty \leqq \alpha\varepsilon \ , \tag{9}$$

and also by Theorem 2.1 (c),

$$||J^* - TJ^*_{N-1}||_\infty = ||J^* - J^*_N||_\infty \leqq \frac{M\alpha^N}{1-\alpha} \ .$$

Then

$$\begin{aligned} ||J^* - TJ||_\infty &\leqq ||J^* - TJ^*_{N-1}||_\infty + ||TJ^*_{N-1} - TJ||_\infty \\ &\leqq \frac{M\alpha^N}{1-\alpha} + \alpha\varepsilon \ . \end{aligned} \tag{10}$$

For the second term in the right-hand side of (8), we use recursively the hypothesis made on function J, together with Inequalities (7) and (9). We obtain

$$\begin{aligned} J &\geqq J^*_{N-1} - \varepsilon \\ &\geqq J^*_N - (\varepsilon + M\alpha^{N-1}) \\ &\geqq (TJ) - (\alpha\varepsilon + \varepsilon + M\alpha^{N-1}) \ , \end{aligned}$$

or, since $TJ = T^f J$:

$$J \ \geqq \ T^f J \ - \ (\alpha\varepsilon + \varepsilon + M\alpha^{N-1}) \ . \tag{11}$$

At this point we can use Lemma 3.1, with $n = 1$ and $C = (\alpha+1)\varepsilon + M\alpha^{N-1}$, to get

$$\begin{aligned} \tilde{U}_N - (TJ) &\leqq ((\alpha+1)\varepsilon + M\alpha^{N-1}) \frac{\alpha}{1-\alpha} \\ &= \frac{\alpha\varepsilon(1+\alpha) + M\alpha^N}{1-\alpha} \ . \end{aligned}$$

Finally,

$$\begin{aligned} \tilde{U}_N - J^* &= \tilde{U}_N - (TJ) + (TJ) - J^* \\ &\leqq \frac{\alpha\varepsilon(1+\alpha)}{1-\alpha} + \frac{M\alpha^N}{1-\alpha} + \frac{M\alpha^N}{1-\alpha} + \alpha\varepsilon \\ &= \frac{2M\alpha^N}{1-\alpha} + \frac{2\alpha\varepsilon}{1-\alpha} \ , \end{aligned}$$

which is the stated bound.

If in addition **Assumption 2** holds, the proof can be made using the inequality

$$J \geqq (T^f J) - (\alpha\varepsilon + \varepsilon)$$

to arrive at:

$$\tilde{U}_N - (TJ) \leqq \frac{\alpha\varepsilon(1+\alpha)}{1-\alpha} , \tag{12}$$

and instead of (11), and the bound becomes

$$||J^* - \tilde{U}_N||_\infty \leqq \frac{M\alpha^N}{1-\alpha} + \frac{2\alpha\varepsilon}{1-\alpha} .$$

□

Theorem 3.4. *Suppose that* **Assumption 1** *holds. Suppose that, for some* N, $||J^*_N - J^*_{N-1}||_\infty \leqq \varepsilon_1$. *Given* J *a bounded function such that* $||J^*_{N-1} - J||_\infty \leqq \varepsilon_2$, *consider a policy* $f \in \Pi_{\text{stat}}$ *such that* $T^f J = TJ$. *Then,*

$$||J^* - \tilde{U}_N||_\infty \leqq \frac{2\alpha(\varepsilon_1 + \varepsilon_2)}{1-\alpha} .$$

Proof. Let us start with the triangular inequality

$$||J^* - \tilde{U}_N||_\infty \leqq ||J^* - J^*_N||_\infty + ||J^*_N - \tilde{U}_N||_\infty .$$

To bound the first term in the r.h.s. of this inequality, observe that, for any $m \in \mathbb{N}$,

$$\begin{aligned} ||J^*_{N+m} - J^*_N||_\infty &\leqq \sum_{k=0}^{m-1} ||J^*_{N+k+1} - J^*_{N+k}||_\infty \\ &= \sum_{k=0}^{m-1} ||T^{k+1}J^*_N - T^{k+1}J^*_{N-1}||_\infty \\ &\leqq \sum_{k=0}^{m-1} \alpha^{k+1}||J^*_N - J^*_{N-1}||_\infty \\ &= \frac{\alpha\ (1-\alpha^m)}{1-\alpha}||J^*_N - J^*_{N-1}||_\infty , \end{aligned}$$

and, taking limits when $m \to \infty$,

$$||J^* - J^*_N||_\infty \leqq \frac{\alpha\varepsilon_1}{1-\alpha} . \tag{13}$$

To bound the second term, we use the facts that $T^f J = TJ$, $T^f \tilde{U}_N = \tilde{U}_N$, and T^f is a contraction of modulus α. With that,

$$\begin{aligned} ||J^*_N - \tilde{U}_N||_\infty &= ||J^*_N - T^f\tilde{U}_N||_\infty \\ &\leqq ||J^*_N - T^f J||_\infty + ||T^f J - T^f\tilde{U}_N||_\infty \\ &= ||TJ^*_{N-1} - TJ||_\infty + \alpha||J - \tilde{U}_N||_\infty \\ &\leqq \alpha\ \varepsilon_1 + \ \alpha\ (||J - J^*_{N-1}||_\infty + ||J^*_{N-1} - J^*_N||_\infty + ||J^*_N - \tilde{U}_N||_\infty) \\ &\leqq \alpha\varepsilon_2 + \alpha(\varepsilon_2 + \varepsilon_1) + \alpha\ ||J^*_N - \tilde{U}_N||_\infty . \end{aligned}$$

which implies

$$(1-\alpha)||J_N^* - \tilde{U}_N||_\infty \leqq 2\alpha\varepsilon_2 + \alpha\varepsilon_1 \ ,$$

or

$$||J_N^* - \tilde{U}_N||_\infty \leqq \frac{2\alpha\varepsilon_2 + \alpha\varepsilon_1}{1-\alpha} \ . \tag{14}$$

Finally, by (13) and (14),

$$||J^* - \tilde{U}_N||_\infty \leqq \frac{2\alpha(\varepsilon_1 + \varepsilon_2)}{1-\alpha} \ .$$

□

Remark 3.1. The statement of Theorem 3.3 can be proved combining Theorems 3.4 and 2.1(c).

In spite of that, we include the proofs of both results, since in any case, they make use of different properties of the dynamic programming operator.

In the proof of Theorem 3.3 the monotonicity property is prevalent. On the other hand, the contractivity property is fundamental in the proof of Theorem 3.4.

Besides, as consequence of Inequality (12) in the proof of Theorem 3.3, with $\varepsilon = 0$, we obtain an order relation, under **Assumption 2**, between the **RH** value U_N and J_N^*, the value of the N-horizon problem, which can not be obtained as a consequence of Theorem 3.4 nor of its proof.

4 Structural Approximations

In the context of stochastic control problems, the controller has to take decisions, based on the knowledge of the current state, but without the certainty of the dynamics of the system, which will be governed by distributions of probabilities on the space of states, known *a priori* for him.

In many situations, the controller may not have the exact information on these probability distributions because they are known only through some statistical method, for example. This lack of information could stem from imprecisions on the measure of quantities involved, and it could be improved by some investment of effort or money. Assessing whether this spending is necessary or excessive is an interesting practical issue. In these cases then, there arises the necessity of having bounds to the errors involved when choosing actions (and then policies) considering the inexact probability distributions. Other model parameters may have also imprecisions. It is also interesting to study the errors produced by such uncertainties.

Through this section, we shall work with approximating stochastic control models of the form

$$\mathcal{M}_n := (\mathcal{S}, \mathcal{A}, \{\mathcal{A}_s : s \in \mathcal{S}\}, \{Q_n^a(\cdot|s) : s \in \mathcal{S}, a \in \mathcal{A}_s\}, r_n, \alpha_n) \ , \tag{15}$$

all differing in the transition probabilities, the reward functions and the discount factors.

In what follows we are interested in approximating the infinite horizon problem defined in (1) through suitable approximating problems.

For the models (15) we define, given a policy $f \in \Pi_{\text{stat}}$, the corresponding dynamic programming operator

$$(T_n^f v)(s) = \sigma\left(r_n^f(s,\cdot) + \alpha_n v(\cdot), s, Q_n^f(\cdot|s)\right) \ ,$$

for $s \in \mathcal{S}$, and we will denote with J_n^f the unique bounded solution (Lemma 2.1) of the equation

$$T_n^f v \ = \ v \ . \tag{16}$$

Let the values of the models $\mathcal{M}_n$ be defined as

$$J_n = \inf_{\pi \in \Pi} J_n^\pi \ .$$

Tidball and Altman in [13], in the context of zero sum games, tried to answer the following questions:

1. Do the sequence of values of the approximating models $\mathcal{M}_n$ converges to the value of the original problem $\mathcal{M}$? If so, is it possible to establish bounds on the error from the approximations?
2. Do optimal (or almost optimal) policies of the approximating models $\mathcal{M}_n$ converge, in some sense, to optimal policies of the limit problem $\mathcal{M}$?
3. Given an optimal policy of the model $\mathcal{M}_n$, will it be an almost optimal for the limit problem if n is sufficiently large?
4. Conversely, given an optimal policy for $\mathcal{M}$, will it be almost optimal for $\mathcal{M}_n$ for all n sufficiently large?

To answer to these questions, the next **Key Theorem** is proved in the same paper ([13, Theorem 2.1., p. 312]). We enunciate here a simpler version, adapted to our case.

Theorem 4.1. Key Theorem
Assume that for the sequence of problems $\{\mathcal{M}_n\}$ *and the original problem* $\mathcal{M}$ *it is verified that*

$$\lim_{n\to\infty} J_n^\pi - J^\pi = 0, \ \textit{uniformly in } \pi \in \Pi \ .$$

Then

(a) $\lim_{n\to\infty} J_n = J^*$.

(b) *For any* $\varepsilon' > \varepsilon$, *there exist* N *such that* π_n^* *is* ε'*-optimal for the limit problem, for* $n \geqq N$.

(c) *Let* π^* *be* ε*-optimal for the limit problem. Then for all* $\varepsilon' > \varepsilon$, *there exist* $N(\varepsilon')$ *such that* π^* *is an* ε'*-optimal policy for all* $n \geqq N(\varepsilon')$.

We can now make precise the sense in which $\mathcal{M}_n$ approximates $\mathcal{M}$.

Assumption 3.

(a) *The sequence of stochastic kernels* Q_n *verifies*

$$\sup_{v:||v||_\infty \leqq 1} |\sigma\left(v, s, Q^a(\cdot|s)\right) - \sigma\left(v, s, Q_n^a(\cdot|s)\right)| \to 0$$

uniformly on $\mathbb{K}$;

(b) *The sequence of discount factors* α_n *verifies* $\alpha_n \to \alpha$;

(c) *The functions* r_n *are bounded on* $\mathbb{K} \times \mathcal{S}$. *That is, there exist constants* $M_n > 0$, *such that, for* $(s, a) \in \mathbb{K}$, $z \in \mathcal{S}$, $|r^a(s, z)| \leqq M_n$;

(d) *The sequence* r_n *satisfies* $r_n \to r$ *uniformly on* $\mathbb{K} \times \mathcal{S}$. *Consequently, the constants* M_n *and* M *can be chosen verifying* $M_n \to M$.

Remark 4.1. In the risk neutral context, the sets $\mathfrak{A}(s,Q)$ reduce to $\{Q\}$ and **Assumption 3** (a) expresses the uniform convergence in total variation norm of the sequence of kernels Q_n to the kernel Q.

The proof of the next lemma is similar to Theorem 2.1 (a). We include that statement for the sake of completeness.

Lemma 4.1. *Under* **Assumption 1** (c), **Assumption 1** (e), *M being the bound on r and* **Assumption 3** (c), *for any stationary policy $f \in \Pi_{\text{stat}}$, the cost in the model $\mathcal{M}$, J^f, defined by (2), and the cost in the models $\mathcal{M}_n$, J_n^f, defined by (16), satisfy:*

$$||J^f||_\infty \ \leqq \ \frac{M}{1-\alpha}, \qquad ||J_n^f||_\infty \ \leqq \ \frac{M_n}{1-\alpha_n} \ .$$

Theorem 4.2. *Consider the models $\mathcal{M}_n$ defined in (15) and $\mathcal{M}$ like in (1). Assume that* **Assumption 1** *holds. Let us define, for any stationary policy $f \in \Pi_{\text{stat}}$, the reward for the limit model $\mathcal{M}$, J^f, by (2), and for the approximating one $\mathcal{M}_n$, J_n^f, by (16). Then, for any $f \in \Pi_{\text{stat}}$:*

$$\begin{aligned} ||J^f - J_n^f||_\infty \ \leqq \ \varepsilon \ := \ & \frac{||r^f - r_n^f||_\infty}{1-\alpha} \\ & + \frac{M}{(1-\alpha)^2} \sup_{s\in\mathcal{S}} \sup_{v:||v||_\infty \leqq 1} |\sigma\left(v,s,Q^f(\cdot|s)\right) - \sigma\left(v,s,Q_n^f(\cdot|s)\right)| \\ & + \frac{M_n}{(1-\alpha_n)(1-\alpha)}|\alpha-\alpha_n|. \end{aligned}$$

and then

$$||J^* - J_n||_\infty \ \leqq \ \varepsilon \ .$$

If **Assumption 3** *holds in addition, then J_n^f converges uniformly to J^f as $n \to \infty$, and the hypotheses of the* **Key Theorem** *hold.*

Proof. If v is a bounded function with C a bound and Q is a transition kernel, in view of the representation (5), $\sigma\left(v,s,Q^{f,g}(\cdot|s)\right) \leqq \sigma\left(C,s,Q^{f,g}(\cdot|s)\right) = C$.

Now, for a fixed stationary policy $f \in \Pi_{\text{stat}}$, for $s \in \mathcal{S}$,

$$|J^f(s) - J_n^f(s)| \ = \ \left|\sigma\left(r^f + \alpha J^f, s, Q^f(\cdot|s)\right) \ - \ \sigma\left(r_n^f + \alpha_n J_n^f, s, Q_n^f(\cdot|s)\right)\right|$$

Introducing the intermediate terms:

$$\begin{aligned} |J^f(s) - J_n^f(s)| \ \leqq \ & |\sigma\left(r^f + \alpha J^f, s, Q^f(\cdot|s)\right) - \sigma\left(r^f + \alpha J^f, s, Q_n^f(\cdot|s)\right)| \\ & + |\sigma\left(r^f + \alpha J^f, s, Q_n^f(\cdot|s)\right) - \sigma\left(r^f + \alpha J_n^f, s, Q_n^f(\cdot|s)\right)| \\ & + |\sigma\left(r^f + \alpha J_n^f, s, Q_n^f(\cdot|s)\right) - \sigma\left(r^f + \alpha_n J_n^f, s, Q_n^f(\cdot|s)\right)| \\ & + |\sigma\left(r^f + \alpha_n J_n^f, s, Q_n^f(\cdot|s)\right) - \sigma\left(r_n^f + \alpha_n J_n^f, s, Q_n^f(\cdot|s)\right)| \\ = \ & \Delta_n^1 + \Delta_n^2 + \Delta_n^3 + \Delta_n^4. \end{aligned}$$

Now we bound each term in the last expression. First, Lemma 4.1 implies:

$$||r^f + \alpha J^f||_\infty \ \leq \ M + \frac{\alpha M}{1-\alpha} \ = \ \frac{M}{1-\alpha} \ .$$

$$\begin{aligned}\Delta_n^1 &\leqq \frac{M}{1-\alpha}\left|\sigma\left(\left(\frac{M}{1-\alpha}\right)^{-1}\left(r^f+\alpha J^f\right),s,Q^f(\cdot|s)\right)\right.\\ &\qquad\left.-\sigma\left(\left(\frac{M}{1-\alpha}\right)^{-1}\left(r^f+\alpha J^f\right),s,Q_n^f(\cdot|s)\right)\right|\\ &\leqq \frac{M}{1-\alpha}\sup_{v:||v||_\infty\leqq 1}\left|\sigma\left(v,s,Q^f(\cdot|s)\right)-\sigma\left(v,s,Q_n^f(\cdot|s)\right)\right|.\end{aligned}$$

Next, observe that $|\sup_\mu f_1(\mu)-\sup_\mu f_2(\mu)| = \max\{\sup_\mu f_1(\mu)-\sup_\mu f_2(\mu), \sup_\mu f_2(\mu)-\sup_\mu f_1(\mu)\} \leq \max\{\sup_\mu(f_1-f_2)(\mu), \sup_\mu(f_2-f_1)(\mu)\} \leq \sup_\mu|(f_1-f_2)(\mu)|$. Accordingly,

$$\begin{aligned}\Delta_n^2 &= \left|\sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}(r^f(z)+\alpha J^f(z))\mathrm{d}\mu(z) \;-\; \sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}(r^f(z)+\alpha J_n^f(z))\mathrm{d}\mu(z)\right|\\ &\leqq \left|\sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}\left[(r^f(z)+\alpha J^f(z))-(r^f(z)+\alpha J_n^f(z))\right]\mathrm{d}\mu(z)\right|\\ &= \alpha\left|\sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}(J^f(z)-J_n^f(z))\mathrm{d}\mu(z)\right|\\ &\leqq \alpha\sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}|J^f(z)-J_n^f(z)|\mathrm{d}\mu(z)\\ &\leqq \alpha||J^f-J_n^f||_\infty.\end{aligned}$$

With the bound given by Lemma 4.1, $||J_n^f||_\infty \leqq \frac{M_n}{1-\alpha_n}$:

$$\begin{aligned}\Delta_n^3 &= \left|\sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}(r^f(z)+\alpha J_n^f(z))\mathrm{d}\mu(z) \;-\; \sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}(r^f(z)+\alpha_n J_n^f(z))\mathrm{d}\mu(z)\right|\\ &\leqq \left|\sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}\left[(r^f(z)+\alpha J_n^f(z))-(r^f(z)+\alpha_n J_n^f(z))\right]\mathrm{d}\mu(z)\right|\\ &= \left|\sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}(\alpha-\alpha_n)J_n^f(z)\mathrm{d}\mu(z)\right|\\ &\leqq |\alpha-\alpha_n|\sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}|J_n^f(z)|\mathrm{d}\mu(z)\\ &\leqq \frac{M_n}{1-\alpha_n}|\alpha-\alpha_n|.\end{aligned}$$

Finally,

$$\begin{aligned}\Delta_n^4 &= \left|\sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}(r^f(z)+\alpha_n J_n^f(z))\mathrm{d}\mu(z) \;-\; \sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}(r_n^f(z)+\alpha_n J_n^f(z))\mathrm{d}\mu(z)\right|\\ &\leqq \left|\sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}\left[(r^f(z)+\alpha_n J_n^f(z))-(r_n^f(z)+\alpha_n J_n^f(z))\right]\mathrm{d}\mu(z)\right|\\ &= \left|\sup_{\mu\in\mathfrak{A}(s,Q_n^f)}\int_{\mathcal{S}}(r^f(z)-r_n^f(z))\mathrm{d}\mu(z)\right|\end{aligned}$$

$$\begin{aligned}
&\leqq \sup_{\mu\in\mathfrak{A}(s,Q_n^f)} \int_{\mathcal{S}} |r^f(z) - r_n^f(z)| \mathrm{d}\mu(z) \\
&\leqq ||r^f - r_n^f||_\infty \ .
\end{aligned}$$

In consequence,

$$\begin{aligned}
||J^f - J_n^f||_\infty \ &\leqq \ ||r^f - r_n^f||_\infty \\
&\quad + \frac{M}{1-\alpha} \sup_{s\in\mathcal{S}} \sup_{v:||v||_\infty\leqq 1} |\sigma\left(v, s, Q^f(\cdot|s)\right) - \sigma\left(v, s, Q_n^f(\cdot|s)\right)| \\
&\quad + \alpha||J^f - J_n^f||_\infty + \frac{M_n}{1-\alpha_n}|\alpha - \alpha_n| \ ,
\end{aligned}$$

which gives the stated bound.

Let us observe that if **Assumption 3** holds, this bound tends to zero uniformly, and so does $J^f - J_n^J$. □

Remark 4.2. If costs $r(\cdot)$ in model $\mathcal{M}$ do not depend on the final state of the transition, then for any f, $\sigma(r^f + g, s, m) = r^f + \sigma(g, s, m)$. The bound for Δ_n^1 is improved as:

$$\begin{aligned}
\Delta_n^1 \ &= \ \left|\sigma\left(r^f + \alpha J^f, s, Q^f(\cdot|s)\right) - \sigma\left(r^f + \alpha J^f, s, Q_n^f(\cdot|s)\right)\right| \\
&= \ \left|\left(r^f + \alpha\sigma\left(J^f, s, Q^f(\cdot|s)\right)\right) - \left(r^f + \alpha\sigma\left(J^f, s, Q_n^f(\cdot|s)\right)\right)\right| \\
&= \ \alpha \ \left|\sigma\left(J^f, s, Q^f(\cdot|s)\right)\right) - \sigma\left(J^f, s, Q_n^f(\cdot|s)\right))\right| \\
&\leqq \ \frac{M\alpha}{1-\alpha}\left|\sigma\left(\left(\frac{M}{1-\alpha}\right)^{-1} J^f, s, Q^f(\cdot|s)\right) - \sigma\left(\left(\frac{M}{1-\alpha}\right)^{-1} J^f, s, Q_n^f(\cdot|s)\right)\right| \\
&\leqq \ \frac{M\alpha}{1-\alpha} \sup_{v:||v||_\infty\leqq 1} |\sigma\left(v, s, Q^f(\cdot|s)\right) - \sigma\left(v, s, Q_n^f(\cdot|s)\right)| \ .
\end{aligned}$$

This improved bound carries over in the bound of Theorem 4.2. Then, in the risk-neutral case (see Remark 4.1), the statement corresponds to the **MDP** version of [6, Theorem 2]. In this risk-neutral case, and when costs do depend on the resulting state, it is standard to transform the problem into one where the cost depends only on the original state, by taking expectations on the future state [9, Chapter 2, p. 20].

5 Example

As an example of application of the **RH** approximation, let us consider a modification of the device maintenance problem presented in [12, Example 1, p. 238].

In this case, the state of a device represents its condition. We shall assume four states, $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$ with s_1 ="faultless" , s_3 ="irreparable and not insured", s_4 ="irreparable and insured" and s_2 an intermediate state between "faultless" and "irreparable".

The actions available at the different states consist on the maintenance strategies and the possibilities of purchase an insurance that reduces the cost of replacing the device when it reaches

the "irreparable" state. More precisely, the set of actions is

$$\begin{aligned}\mathcal{A} = \quad \{ \quad & a_1 = \text{"do not make maintenance tasks and do not buy insurance"},\\ & a_2 = \text{"perform maintenance and do not buy insurance"},\\ & a_3 = \text{"do not make maintenance tasks and buy insurance"},\\ & a_4 = \text{"perform maintenance and buy insurance"} ,\\ & a_5 = \text{"replace item"} \quad \} .\end{aligned}$$

According to the model, since at the state s_1 the action "replace" is meaningful and at states s_3, s_4 "perform maintenance" and "do not make maintenance tasks" are not options, the sets of corresponding available actions are

$$\mathcal{A}_{s_1} = \mathcal{A} \setminus \{a_5\},\ \mathcal{A}_{s_2} = \mathcal{A},\ \mathcal{A}_{s_3} = A_{s_4} = \{a_5\} .$$

The non-zero transition probabilities and costs are given by

$$\begin{aligned}&Q^{a_1}(s_1|s_1) = Q^{a_3}(s_1|s_1) = Q^{a_1}(s_3|s_1) = Q^{a_3}(s_4|s_1) = 0.1\\ &Q^{a_1}(s_2|s_1) = Q^{a_3}(s_2|s_1) = 0.8 , \quad Q^{a_1}(s_2|s_2) = Q^{a_3}(s_2|s_2) = 0.6\\ &Q^{a_1}(s_3|s_2) = Q^{a_3}(s_4|s_2) = 0.4 , \quad Q^{a_2}(s_1|s_1) = Q^{a_4}(s_1|s_1) = 0.25\\ &Q^{a_2}(s_2|s_1) = Q^{a_4}(s_2|s_1) = Q^{a_2}(s_2|s_2) = Q^{a_4}(s_2|s_2) = 0.7\\ &Q^{a_2}(s_1|s_2) = Q^{a_4}(s_1|s_2) = 0.1 , \quad Q^{a_2}(s_3|s_1) = Q^{a_4}(s_4|s_1) = 0.05\\ &Q^{a_2}(s_3|s_2) = Q^{a_4}(s_4|s_2) = 0.2 , \quad Q^{a_5}(s_1|s_3) = Q^{a_5}(s_1|s_4) = 1 ,\end{aligned}$$

$$\begin{aligned}&r^{a_2}(s_1) = r^{a_2}(s_2) = 10 , \quad r^{a_3}(s_1) = 20 , \quad r^{a_4}(s_1) = 30\\ &r^{a_3}(s_2) = 40 , \quad r^{a_4}(s_2) = 55 , \quad r^{a_5}(s_3) = 200 , \quad r^{a_5}(s_4) = 100 .\end{aligned}$$

In this example we evaluate the performance of strategies through the mean deviation measure and using the expression of σ given in [12, Example 4, Equation (16), p. 244]:

$$\sigma(v, x, m) \ = \ < v, m > \ + \ \kappa(x) \left(< ((v - < v, m >)_+)^2, m > \right)^{1/2} .$$

where $\kappa \in [0, 1]$ is a parameter modeling the risk attitude of the controller, and $(y)_+$ denotes $\max\{y, 0\}$. With it, the dynamic programming operator takes the form:

$$\begin{aligned}(Tv)(s) \quad = \quad \min_{a \in \mathcal{A}_s} \Bigg\{ & r^a(s)\\ & + \alpha \left(\sum_{z \in \mathcal{S}} v(z) Q^a(z|s) + \kappa \left[\sum_{w \in \mathcal{S}} \left[\left(v(w) - \sum_{z \in \mathcal{S}} v(z) Q^a(z|s) \right)_+ \right]^2 Q^a(w|s) \right]^{\frac{1}{2}} \right) \Bigg\}\end{aligned}$$

By application of the **RH** procedure with horizon 100 to the problem with discount $\alpha = 0.9$, we obtain four different strategies, depending on the risk parameter κ (assumed to be the same for all states).

For $\kappa = 0$ and $\kappa = 0.046$, the strategy obtained is (a_1, a_2, a_5, a_5), consisting on not doing anything when the device is new, performing maintenance in the intermediate state, and not buying insurance in any case.

For $\kappa = 0.047$, $\kappa = 0.562$, the strategy is (a_2, a_2, a_5, a_5), consisting in performing the maintenance at the new and the intermediate state, and again not buying the insurance.

For $\kappa = 0.563$, $\kappa = 0.567$, the strategy is (a_2, a_3, a_5, a_5), which states: performing maintenance without buying the insurance in new devices, and buying the insurance at the intermediate state, without performing any maintenance.

Finally, for $\kappa = 0.568$, $\kappa = 1$, the **RH** strategy is (a_3, a_3, a_5, a_5): do not do anything but buy an insurance, when the state is not irreparable.

We calculate the error bounds in the **RH** procedure provided by Theorems 3.3 (for positive costs, applied with $\varepsilon = 0$) and 3.4 (with $\varepsilon_2 = 0$):

$$||J^* - J_N||_\infty \leqq B1 := \frac{200 \times 0.9^N}{1 - 0.9} \ , \quad ||J^* - J_N||_\infty \leqq B2 := \frac{2 \times 0.9 \times ||J_N^* - J_{N-1}^*||_\infty}{1 - 0.9} \ ,$$

as well as the relative error bounds

$$RB1 = \frac{B_1}{\min J^*} \ , \quad RB2 = \frac{B_2}{\min J^*}$$

in this example for horizons $N = 20$, $N = 35$ and $N = 50$ and for the risk parameters $\kappa = 0.01$, $\kappa = 0.5$ and $\kappa = 0.99$.

$\kappa = 0.01$	$N = 20$	$N = 35$	$N = 50$
$B1$	243.153	50.063	10.308
$B2$	110.411	22.733	4.680
$RB1$	0.757	0.156	0.032
$RB2$	0.344	0.071	0.015

Table 1: Error bounds for parameter $\kappa = 0.01$. $\min J^* = 321.345$.

$\kappa = 0.5$	$N = 20$	$N = 35$	$N = 50$
$B1$	243.153	50.063	10.308
$B2$	158.974	32.731	6.739
$RB1$	0.513	0.106	0.022
$RB2$	0.335	0.069	0.014

Table 2: Error bounds for parameter $\kappa = 0.5$. $\min J^* = 474.346$.

$\kappa = 0.99$	$N = 20$	$N = 35$	$N = 50$
$B1$	243.153	50.063	10.308
$B2$	167.795	34.538	7.111
$RB1$	0.476	0.098	0.020
$RB2$	0.328	0.068	0.014

Table 3: Error bounds for parameter $\kappa = 0.99$. $\min J^* = 511.081$.

Since the bound given by Theorem 3.3 does not depend on the parameter κ, values $B1$ coincides for the same values of N, but is not the case for $RB1$, which depends on the optimal value J^*.

In this instance of the problem, the **RH** policies converge to the optimal ones after 3 iterations for the case $\kappa = 0.01$ and after 4 iterations for the values $\kappa = 0.5$ and $\kappa = 0.99$.

From these experiments, we draw some preliminary conclusions for approximate rolling horizon:

- The numerical values obtained from Theorems 3.3 and 3.4 are disappointingly large, however the relative values are more reasonable.

  The bound obtained from Theorem 3.4 is better: this illustrates the benefit of using successive steps of Value Iteration instead of only the last one.

- As in standard risk-neutral dynamic programming, the policies appear to converge much faster than values, and values much faster than what bounds predict.

  The first fact validates in practice the use of Rolling Horizon with a relatively small horizon. The second one suggests that developing bounds that take into account some of the problem structure is an interesting challenge.

Concerning structural approximations, the practical issue is currently the way to evaluate numerically (or bound efficiently) the "distance" between two transition kernels:

$$\sup_{v:||v||_\infty \leqq 1} \left| \sigma\left(v, s, Q^f(\cdot|s)\right) - \sigma\left(v, s, Q_n^f(\cdot|s)\right) \right| \ .$$

6 Concluding Remarks

Through this work we have dealt with risk-averse **MDP** with discounted cost, studying the performance of two different ways of approximations of value functions and optimal policies.

In Section 3, we have studied the performance of the Rolling Horizon procedure and of an Approximate Rolling Horizon procedure, the former being a particular case of the latter. We have proved the uniform geometrical convergence of the values related to the **RH** procedure to the optimal cost function. For the **ARH** we have obtained the same convergence bound plus a constant term in the horizon, which can be improved with the initial approximation.

In Section 4 we have studied approximations of the value function of the infinite-horizon risk-averse **MDP** and their optimal policies, by considering it as a limit of a sequence of approximating models. The approximating models are obtained by perturbing transition probabilities, cost functions and discount factors. The perturbation is measured by supremum norms for functions and parameters, and we introduce a metric for perturbation of operators. We obtain a bound in terms of these perturbations, which implies the uniform convergence of values when parameters are converging uniformly to the original ones.

Finally, in Section 5 we have provided a numerical example and raised some issues for future research.

Acknowledgements

E. Della Vecchia wishes to thank the help and the encouragements of Andrzej Ruszczyński and Darinka Dentcheva during the Workshop of Stochastic Optimization organized at the Universidad de Chile in December 2012.

References

[1] Altman, E.; *Constrained Markov Decision processes*, Chapman and Hall, 1999.

[2] Cavazos-Cadena, R.; *Finite-state approximations for denumerable state discounted Markov decision processes.* Applied Mathematics and Optimization, 14, 1, 1986, pp. 1–26.

[3] Çavuş, O., Ruszczyński, A.; *Risk-Averse Control of Undiscounted Transient Markov Models.*, manuscript, 2013.

[4] Chang H., Marcus, S.; *Two-person zero-sum games: receding horizon approach.* IEEE Transactions on Automatic Control, 48, 11, 2003.

[5] Della Vecchia E., Di Marco S., Jean-Marie A., *Rolling horizon procedures in Semi-Markov Games: The Discounted Case.* INRIA Research Report 8019, July 2012. http://hal.inria.fr/hal-00720351

[6] Della Vecchia E., Di Marco S., Jean-Marie A., *Structural approximations in discounted Semi-Markov Games.* INRIA Research Report 8162, November 2012. http://hal.inria.fr/hal-00764217

[7] Hernández-Lerma O., *Adaptive Markov Control Processes.* Springer-Verlag 1989.

[8] Hernández-Lerma O, Lasserre J.B.; *Error bounds for rolling horizon policies in discrete-time Markov control processes.* IEEE Transactions on Automatic Control, 35, 10, 1990, pp. 1118–1124

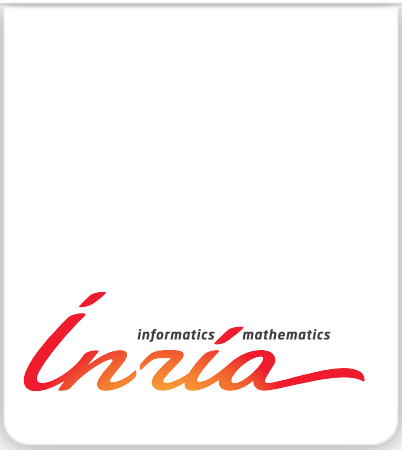
[9] Puterman L., *Markov Decision Processes.* Wiley and Sons, 2005.

[10] Ruszczyński, A., Shapiro, A.; *Conditional risk mappings.* Math. Oper. Res. 31, 2006, pp. 544 - 561.

[11] Ruszczyński, A., Shapiro, A.; *Optimization of Convex Risk Functions.* Math. Oper. Res. 31, 2006, pp. 433–452.

[12] Ruszczyński, A.; *Risk-averse dynamic programming for Markov decision processes.* Math. Programming, Series B, 125, 2010, pp. 235–261.

[13] Tidball M., Altman E., *Approximations in Dynamic Zero-Sum Games, I* SIAM Journal on Control and Optimization, 1993, pp. 311–328.



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399